UNIVERSIDAD DEL BÍO-BÍO

TAREA 1 | COCHILCO 2014–2026

Producción mensual de cobre de mina en Chile

Análisis exploratorio, limpieza y visualización de datos

Ciencia de Datos para Economía y Gestión

Dataset oficial de producción minera

1. Objetivo, fuente y unidad

Pregunta de trabajo

¿Qué revela la producción mensual de cobre de mina en Chile cuando se limpia y reorganiza el archivo original de COCHILCO?

Dataset oficial

Producción de Cobre Mina por Empresa Mensual, en miles de toneladas métricas de cobre fino.

Período limpio

Enero 2014 a marzo 2026

Unidad

Miles de T.M. de cobre fino

Entrega

Notebook + datos limpios + presentación

2. Variables y formatos usados

Dos formas de mirar la misma base

Formato	Uso principal
Ancho	Una fila por mes y una columna por faena. Sirve para describir el total nacional, correlaciones y series mensuales.
Largo	Una fila por mes y faena. Sirve para comparar faenas, calcular rankings y ver concentración productiva.

Variables centrales

Variable	Rol en el análisis
total_chile	Producción nacional
total_codelco	Agregado institucional
escondida	Faena privada mayor
collahuasi	Faena privada relevante
los_pelambres	Faena privada relevante

El formato largo no inventa datos: solo reordena las columnas de faenas en filas para poder compararlas.

3. Del Excel bruto al dataset limpio

Excel COCHILCO con títulos y notas → Filas mensuales identificadas → Tipos corregidos fecha + números → Datos limpios ancho y largo

Problemas encontrados

- Encabezados institucionales.
- Resúmenes anuales mezclados.
- Meses futuros sin dato informado.
- Nombres de columnas no estandarizados.

Decisiones aplicadas

- Eliminar filas no mensuales.
- Filtrar meses con total nacional igual a cero.
- Convertir variables a tipos correctos.
- Crear formato largo para analizar faenas.

4. Limpieza cuantificada

Cambios de estructura

Etapa	Filas	Cols.	Criterio
Excel leído	171	44	Incluye títulos, años y fuente.
Sin filas vacías	170	44	Se eliminan filas totalmente vacías.
Solo meses	156	44	Se conservan fechas mensuales.
Total Chile > 0	147	44	Se dejan meses con dato informado.
Formato largo	5.694	5	Fecha, faena y producción positiva.

Acciones implementadas

- ▷ Filtrado de filas no mensuales.
- ▷ Conversión de fechas y columnas numéricas.
- ▷ Eliminación de meses sin producción nacional informada.
- ▷ Transformación a formato largo.

5. ETL: extracción

```
archivo = DATOS / "produccion_cobre_mina_empresa_mensual.xlsx"

df_raw = pd.read_excel(
    archivo,
    sheet_name=0,
    header=6,
)

df_raw.shape
```

Qué hace esta fase

Se toma el Excel original de COCHILCO y se carga en Python. No se corrige nada todavía: el objetivo es conservar una copia bruta para revisar cómo venía el archivo.

El parámetro `header=6` indica que los nombres de columnas reales comienzan en esa fila del Excel.

6. ETL: transformación

```
df = df_raw.copy()
df = df.dropna(how="all")

columnas_limpias = {
    col: limpiar_nombre_columna(col)
    for col in df.columns
}

df = df.rename(columns=columnas_limpias)
df = df.rename(columns={df.columns[0]: "fecha"})
```

Limpieza estructural

Se trabaja sobre una copia, se eliminan filas totalmente vacías y se estandarizan los nombres de columnas para que sean fáciles de usar en código.

Ejemplo: una columna con espacios, mayúsculas o acentos queda convertida a un nombre simple como `total_chile`.

7. ETL: transformación analítica

```python
df = df[df["total_chile"].fillna(0) > 0].copy()

df_largo = df.melt(
    id_vars="fecha",
    value_vars=columnas_empresas,
    var_name="empresa_faena",
    value_name="produccion_miles_tm_cobre_fino",
)

df_largo = df_largo.dropna(
    subset=["produccion_miles_tm_cobre_fino"]
)
df_largo = df_largo[
    df_largo["produccion_miles_tm_cobre_fino"] > 0
].copy()
```

Preparación para analizar

Se dejan solo meses con producción nacional informada y se crea el formato largo: una fila por mes y por faena.

Esta transformación permite comparar faenas, calcular rankings y hacer histogramas por empresa/-faena.

8. ETL: carga y resultados

```python
df.to_csv(
    DATOS_LIMPIOS / "produccion_cobre_mensual_limpio_ancho.csv",
    index=False,
)

df_largo.to_csv(
    DATOS_LIMPIOS / "produccion_cobre_mensual_limpio_largo.csv",
    index=False,
)

resumen_empresas.to_csv(
    DATOS_LIMPIOS / "resumen_produccion_por_empresa.csv"
)
```

Qué se entrega

La fase de carga guarda los productos finales: dataset limpio en formato ancho, dataset largo y resumen por faena.

Estos archivos son la base para los gráficos, tablas, notebook ejecutado, informe y presentación.

9. Radiografía del dataset

Formato ancho
147 meses

Variables
44 columnas

Formato largo
5.694 registros

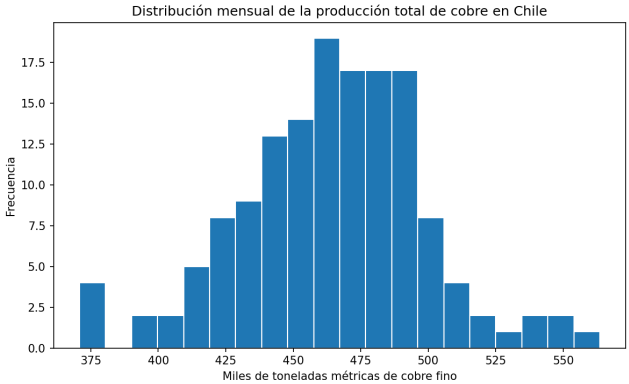
Estadísticos descriptivos

Variable	Media	Mín.	Máx.
Total Chile	464,5	370,8	563,6
Total Codelco	129,6	79,9	177,7
Escondida	95,0	0,0	133,6
Collahuasi	44,5	17,0	60,3
Los Pelambres	29,0	14,0	38,1

Lectura principal

La serie nacional se mueve en un rango relativamente estable, pero algunos meses aparecen fuera del comportamiento habitual y deben revisarse como posibles eventos económicos u operacionales.

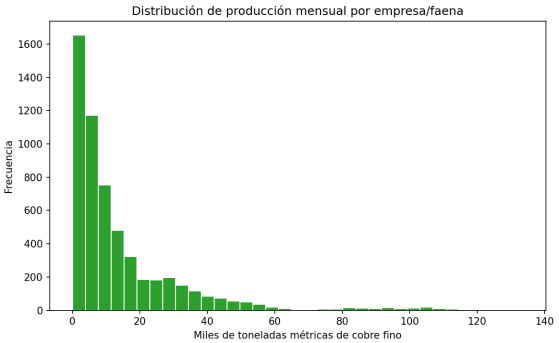
10. Distribución de la producción nacional



Por qué histograma
La variable es numérica continua y permite observar frecuencia, dispersión y valores inusualmente altos o bajos.

Resultado
La mayor parte de los meses se concentra cerca de 445 a 487 mil toneladas.

11. Concentración por empresa y faena

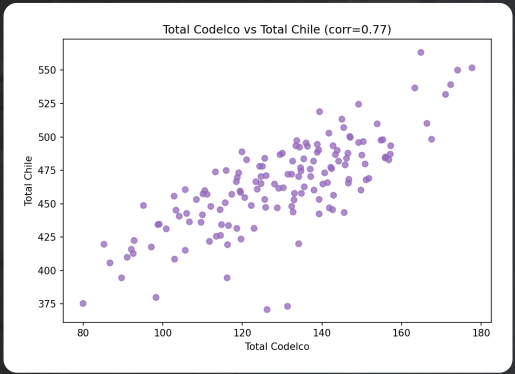


Faenas con mayor promedio mensual positivo

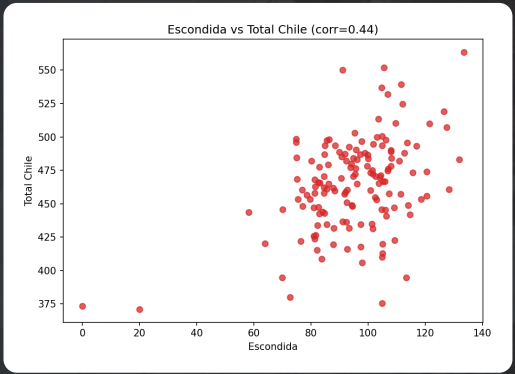
Faena	Prom.
Escondida	95,6
Chuqui y R. Tomic	51,3
Collahuasi	44,5
El Teniente	35,2
Los Pelambres	29,0

Promedio calculado solo en meses con producción positiva. Pocas operaciones grandes explican una parte importante del volumen nacional.

12. Relaciones con la producción nacional



Correlación Codelco–Chile: 0,77



Correlación Escondida–Chile: 0,44

Codelco, como agregado, se mueve más cerca del total nacional que una faena individual, aunque Escondida sea la operación de mayor producción promedio positiva.

13. Correlaciones relevantes

Mayor relación con total_chile

Variable	Corr.	Intensidad
total_codelco	0,77	Alta
chuqui_y_r_tomic	0,75	Alta
chuquicamata	0,71	Alta
el_teniente	0,59	Media
los_pelambres	0,57	Media
escondida	0,44	Media
collahuasi	0,30	Baja

Lectura

La relación más fuerte aparece en Codelco porque agrupa varias divisiones. En cambio, una faena individual puede ser grande, pero no necesariamente explica sola el movimiento nacional mensual.

Este punto justifica el uso de scatter plots: ambas variables son numéricas y permiten observar asociación, dispersión y meses alejados del patrón.

14. Valores faltantes y meses sin dato

Antes del filtro final

Variable	Faltantes	Porcentaje
chuquicamata	9	5,77 %
mantoverde	9	5,77 %
los_pelambres	9	5,77 %
zaldivar	9	5,77 %
centinela_sulfuros	9	5,77 %
quebrada_blanca	9	5,77 %

Interpretación

Los faltantes se concentran en meses donde el archivo todavía no informa producción nacional completa. Por eso no se imputan: se filtran los meses con total_chile igual a cero.

Después de limpiar, el dataset ancho queda con 147 meses, 44 columnas y sin filas duplicadas. El período final usado es enero 2014 a marzo 2026.

15. Outliers: no todo valor extremo es error

Meses detectados por rango intercuartil

Mes	Total Chile
Feb. 2017	370,8
Mar. 2017	373,1
Dic. 2018	550,1
Dic. 2019	551,8
Feb. 2023	380,0
Dic. 2024	563,6
Feb. 2026	375,4

Decisión

Se conservaron estos meses porque pueden representar mantenciones, paralizaciones, cambios operacionales o cierres de periodo. En minería, un outlier puede ser información, no ruido.

16. Asimetría y concentración

Medidas de forma

Variable	Asimetría	Curtosis
total_chile	-0,12	0,69
total_codelco	-0,13	-0,29
escondida	-1,53	7,48
collahuasi	-0,36	0,48
los_pelambres	-0,76	0,91
el_teniente	-0,72	0,57

Resultado clave

Escondida destaca por asimetría negativa y curtosis alta. Esto indica meses especialmente bajos y presencia de valores extremos, coherente con una faena de gran escala que puede alterar el total nacional cuando se desvía de su nivel habitual.

La decisión fue no imputar ni borrar automáticamente estos valores, sino tratarlos como observaciones relevantes del proceso productivo.

17. Conclusiones

Hallazgo central

La producción mensual chilena de cobre es relativamente estable, pero está concentrada en pocas operaciones grandes y presenta meses puntuales que merecen explicación económica u operacional.

- Escondida lidera por producción promedio positiva.
- Codelco explica mejor el movimiento del total nacional.
- La limpieza fue necesaria por la estructura original del Excel.
- El formato largo deja el dataset listo para análisis posteriores.

18. Fuentes y créditos visuales

- Datos: COCHILCO, "Producción de Cobre Mina por Empresa Mensual".
- Chuquicamata: Owen Cliffe / Zootalures, Wikimedia Commons, CC BY-SA 3.0.
- Escondida: NASA Johnson Space Center / Expedition 22 crew, Wikimedia Commons, dominio público en Estados Unidos.
- Cátodo de cobre: WeHaKa, Wikimedia Commons, CC BY-SA 4.0.
- Mineral de cobre: S. Rae, Wikimedia Commons, CC BY 2.0.
- Camión minero: Phil Whitehouse, Wikimedia Commons, CC BY 2.0.
- Gráficos: elaboración propia en Python a partir de COCHILCO.

Gracias